

District of Columbia Interagency Data Team

October 21, 2021

2:00 PM – 3:30 PM

Office of Chief Technology Officer

Meeting Hosted via WebEx Events

Agenda

- **Welcome Remarks with News & Updates**
Barney Krucoff, Chief Data Officer, Office of the Chief Technology Officer
- **DC Health Disease Surveillance Data Engine**
Cindy Knoll, Accenture Engagement Lead, DC Health
Johnny Nguyen, Accenture Data Engine Technical Architect and Application Lead, DC Health
- **Data Lake Upgrade**
Matthew Sokol, Big Data & Data Integration PM, Office of the Chief Technology Officer
- **Data Report**
Matthew Crossett, Enterprise Mapping Project Manager, Office of the Chief Technology Officer

Visit <https://octo.dc.gov/page/interagency-data-team> for presentation and minutes.

Meeting Notes

Best attempt to capture notable comments and questions from attendees (paraphrased). For complete presentation visit: <https://octo.dc.gov/page/interagency-data-team>

Presenter: Barney Krucoff (OCTO) – Systems & Training News

- Tableau: review of current production and test environment versions
 - Planned upgrades scheduled for 10/28/2021. Check the Tableau community Teams site for more information including outages.
 - Join the [Tableau Community Teams site here](#)
 - Creators should save all work and log off by 6 PM Thursday October 28, 2021.
 - Tableau version 2020.2.3 allows for multiple map layers and spatial area calculations
- Tableau Conference: November 9-11, 2021. Free for DC.gov attendees. Register at <https://www.tableau.com/events/conference/schedule>
- Tableau site administrator's meeting tentatively scheduled for November 5, 2021.
- MicroStrategy: review current production and test environment versions.
 - Planned upgrades to production server on 10/22/2021
 - Planned upgrades to lower environment on 10/21/2021

- Next MicroStrategy stakeholder’s meeting scheduled for November 4, 2021
- Big Data and Data Integration
 - Cloudera platform phase 1 has been upgraded with enhanced security, easier Kerberos authentication, apache spark 2.x – completed on 10/9/2021.
 - Phase 2 upgrade is TBD
 - Data.in.dc.gov now updated with the latest 2021 Enterprise Dataset Inventory
 - API gateway upgrade from version 9.4 to 10 in second quarter. There will be additional API management tools
- Geographic Information Systems
 - Esri GeoEvent is now integrated with Data Lake.
 - ArcGIS Portal currently at 10.8.1
 - New updated server hardware for ArcGIS Servers that support Maps2. Resulting from DC.gov data center migration.
 - 2021 Orthophoto will be available first week in November 2021.
 - Planimetric data release expected at the end of November 2021. Currently under quality control
- REMINDER: 2022 Enterprise Dataset Inventory is open for editing as of 9/16/2021.
 - Agency Data Officer meeting scheduled for November 3, 2021.
 - EDI will be closed to editing mid-December
- User Groups
 - GIS – scheduled for October 28, 2021
 - Data Science – TBD.
- GIS and BI Training – please check <http://octo.in.dc.gov/page/gis-and-bi-apps-and-services> for posted dates and registration.

Presenter: Cindy Knoll, Johnny Nguyen & Andersen Andrews (DC Health) – Disease Surveillance Data Engine

- Future state Disease Surveillance combines agility, flexibility, scalability, and automation. The increased automation involves building a highway into the Data Engine, which will drive enhanced reporting capabilities, result in time savings and help facilitate more informed health policy decisions.
- Data Engine automated the work of many workers who were doing the analysis manually at the beginning of the pandemic
- After almost 2 years of COVID-19, DC Health,
 - Experienced a massive uptick in electronic lab records due to COVID-19 data
 - The agency systems needed to be upgraded for them to handle the increase in data. Needed a contact tracing environment. Needed a disease surveillance environment
- In coordination with OCTO and Accenture, DC Health was able to scale and automate the data to build the “Data Engine” which automated report capabilities through a data warehouse.
- This saved time through data ingestion, cleansing and de-duplication to produce deliverables for health policy reports and decisions.
- The Data Engine uses Informatica MDM for data ingestion, cleansing and de-duplication
- Key data sources
 - Lab test results

- Vaccination data
- Historical case investigations
- COVID-19 test results
- Reports
- Patient information
- Birth and death certificates
- The Data Engine creates a relational model to bring together data about each person.
- The MDM Informatica rules determine whether there is an existing record for an individual or whether the individual is not in the system and a new record needs to be created. Most of the data sources use the District's [Master Address Repository API](#) to geocode patient addresses.
- MDM Match Rules to identify if same person or not:
 - Rule 1: Data of birth, first and last name
 - Rule 2: DOB, phone, person (fuzzy)
 - Rule 3: DOB, address, person name (fuzzy)
 - Manual merge
- The high level architecture of the Data Engine is as follows. See presentation for more clarification.
 - Identify the source data attributes
 - Create the table structures
 - Load the data
 - Match/merge the data for unique person
 - Publish the data to the consumption layer
- OCTO supports many of the components to help Accenture build the Data Engine,
 - Master Address Repository
 - Tableau Server
 - Informatica MDM
 - Oracle DB
- Other remarks,
 - DC Health noticed that there were lots of databases operating in silos.
 - The mandate of DC Health is to make policies based on data.
 - DC Health needs to use the investments that OCTO had already done.
 - Engaged Accenture because they were one of the first groups to build a contact tracing platform
 - Objective: build a common environment, needed the staff and skill in place, and looked to OCTO and Accenture
- *Question:* will DC Health continue to add to the Data Engine architecture?
 - Yes, because of our Agile approach to the project.
- *Question:* Have you tun into issues with the fuzzy matching when it comes to twins, triplets and siblings?
 - Yes. There is fuzzy matching on the first name in the case of twins, to create tie breakers. For Example, say you have twins named "Daniel" and "Danielle". The Informatica MDM uses a database of similar names to create scores of likenesses between names. In the cases of not 100% matches, then need to manually review.

Presenter: Matthew Sokol (OCTO) – Data Lake Upgrade

- Key Focus Points
 - Cloudera platform updated to CDP 7.1.7

- Includes better security with Apache Ranger – stronger, tag based security
- Improved data governance and auditing
- Better data processing with Apache Spark 2.x. Opportunity to upgrade to 3.x
- Post Upgrade
 - Kerberos change is automated so you don't have to replace ticket
 - Expansion of tools to connect to Data Lake with R Studio, etc
 - Data anonymization with integration of Kogni. For example, masking SSN
 - Project support – now on-boarding more data projects and machine learning

Presenter: Matthew Crossett (OCTO) – Data Report

New Datasets in Open Data DC since last meeting September 16, 2021

- Comprehensive Plan 2021
 - Generalized Plan Policy
 - Future Analysis Areas
 - Resilience Analysis Areas
 - Future Land Use
 - State of Washington DC in 2021
- Traffic Signals
- Brain Health Initiative
 - Advanced Care and End of Life Planning
 - Community-Based Dementia Care
 - Case Management Care Planning and Care Navigation
 - Connection to Services Helplines and Crisis Care
 - Food and Nutrition
 - Healthcare for Homebound Older Adults and End of Life Care
 - Home-Based Homemaker Services
 - Home-Based Personal Care Aides
 - Individual and Home Safety
 - Medication Support
 - Money Management and Financial Literacy
 - Non-Emergency Transportation
 - Residential Long Term Memory Care
 - Social Connection Engagement and Networking
 - Brain Health Web-Based Resource

Updated Datasets in Open Data DC since last meeting March 9, 2021

- Child Development Centers
- Camera Enforcement Locations
- Moving Violation August and Sept
- Parking Violation August and Sept
- Roadway Alleys
- Roadway Driveways
- Roadway Major Roads
- Roadway Ramps
- Roadway Service Roads
- Snow Removal Areas and Zones

- Wetland Inventory

END